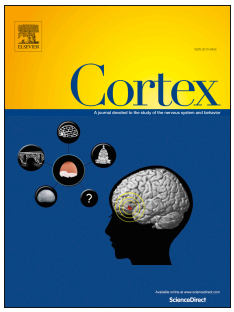
Accepted Manuscript



Brain-decoding fMRI reveals how wholes relate to the sum of parts

Jonas Kubilius, Annelies Baeck, Johan Wagemans, Hans P. Op de Beeck

PII: S0010-9452(15)00052-0

DOI: 10.1016/j.cortex.2015.01.020

Reference: CORTEX 1389

To appear in: *Cortex*

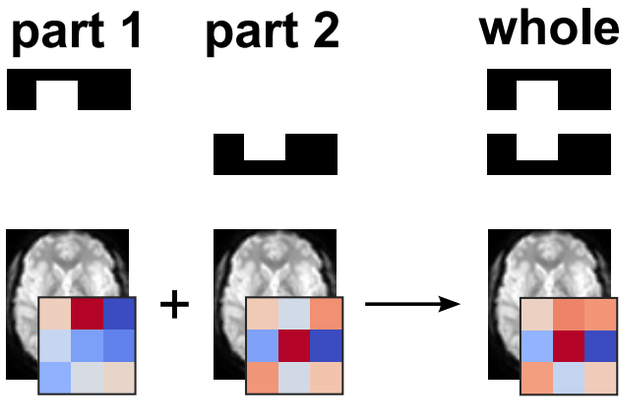
Received Date: 28 July 2014

Revised Date: 3 December 2014

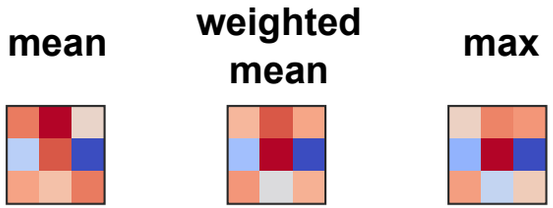
Accepted Date: 27 January 2015

Please cite this article as: Kubilius J, Baeck A, Wagemans J, Op de Beeck HP, Brain-decoding fMRI reveals how wholes relate to the sum of parts, *CORTEX* (2015), doi: 10.1016/j.cortex.2015.01.020.

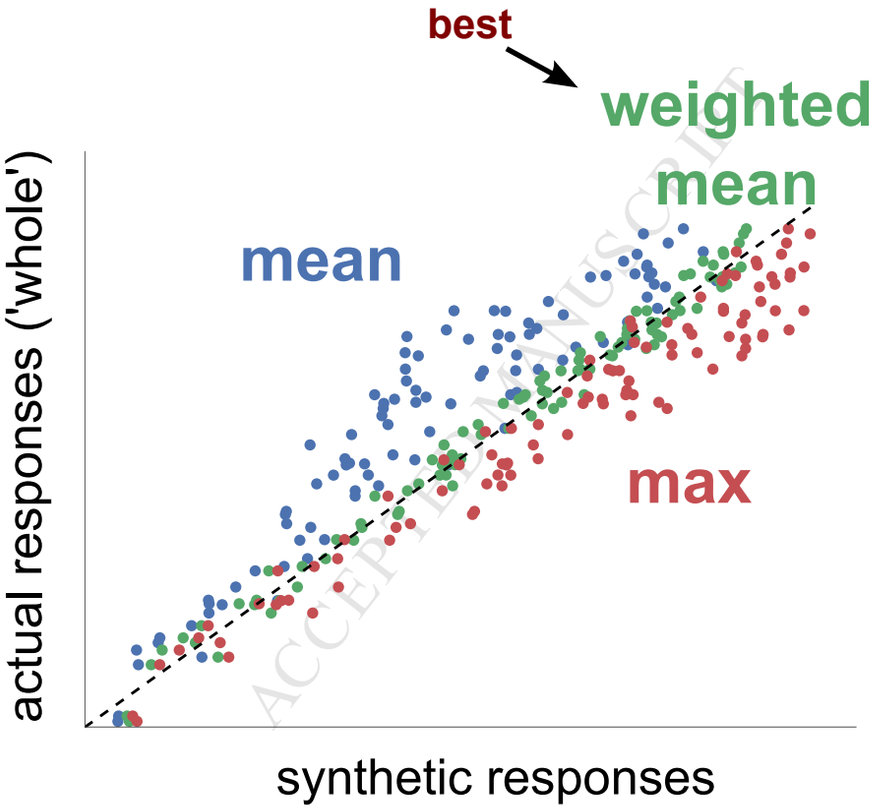
1 record fMRI patterns of response



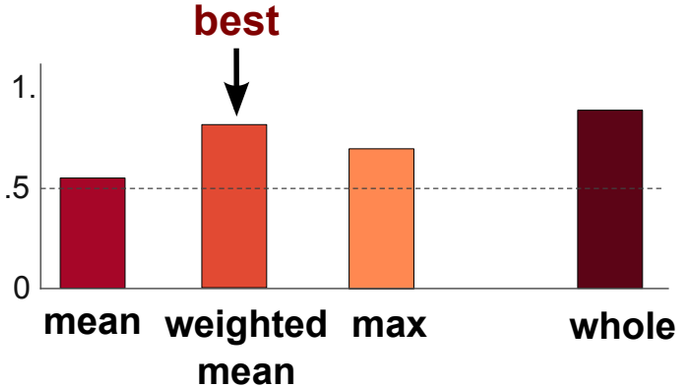
2 combine parts using:



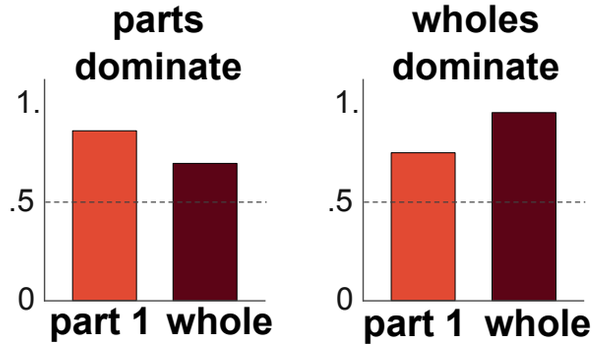
3 responses in a single voxel



4 multi-voxel: decoding synthetic vs. actual



5 multi-voxel: decoding parts vs. wholes



Brain-decoding fMRI reveals how wholes relate to the sum of parts

Jonas Kubilius[1,2], Annelies Baeck[1,2], Johan Wagemans[2], Hans P. Op de Beeck[1,*]

[1]Laboratory of Biological Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[2]Laboratory of Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

***Correspondence:** Hans P. Op de Beeck, Laboratories of Biological and Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102 bus 3714, B-3000 Leuven, Belgium. Email: Hans.OpdeBeeck@ppw.kuleuven.be. Phone: +32 16 326 039.

Keywords: functional magnetic resonance imaging, multi-voxel pattern analysis, nonlinear.

Abstract

The human brain performs many nonlinear operations in order to extract relevant information from local inputs. How can we observe and quantify these effects within and across large patches of cortex? In this paper, we discuss the application of multi-voxel pattern analysis (MVPA) in functional magnetic resonance imaging to this issue. Specifically, we show how MVPA (i) allows to compare various possibilities of part combinations into wholes, such as taking the mean, weighted mean, or the maximum of responses to the parts; (ii) can be used to quantify the parameters of these combinations; and (iii) can be applied in various experimental paradigms. Through these procedures fMRI helps to obtain a computational understanding of how local information is integrated into larger wholes in various cortical regions.

1. Introduction

Information processing in the human brain is highly nonlinear. Many famous visual illusions illustrate this idea particularly well. Consider, for instance, a modified Kanizsa display (Fig. 1a, 'whole'; von der Heydt et al., 1984). When the two black rectangles with notches in them are properly aligned, an illusory white rectangle emerges (especially if the white rectangle is moving) – an effect that is hard to predict from the parts alone. Other striking illustrations come from the drawings of artist M.C. Escher in which the figure/ground segregation can be completely reversed (Fig. 1b). When one set of elements is seen as figure, the others seem to lose their shape. Such nonlinear effects are also common in higher-level visual processing. One example is the composite face effect (Young et al., 1987) where the perception of face parts is affected by the context of a whole face (Fig. 1c).

But how exactly do these nonlinear mechanisms operate? How do we combine parts into wholes and what criteria must be met? After all, not all processing is nonlinear. It suffices to flip parts in Fig. 1a and the combination does no longer produce the illusory rectangle, that is, the combination is now a simple sum of the two parts. Understanding the principles of neural computations governing nonlinear integration is potentially key in understanding the entire visual processing. Unsurprisingly, this question has been central to vision research from its early days, as emphasized by the rise of Gestalt psychology. Gestalt psychologists recognized that in many cases the wholes appear to be not equal to the sum of their parts, and attempted to come up with a list a laws, such as grouping by proximity or good continuation, that would describe, at least in qualitative terms, under what circumstances these nonlinear effects emerge (for a review, see Wagemans et al., 2012).

These early Gestaltist ideas have further gained strong support from multiple neurophysiological studies. In their seminal experiment, von der Heydt and colleagues (1984) used a modified Kanizsa display consisting of two black bars with notches in them, as depicted in Fig. 1a, to investigate neural responses in monkey visual areas V1 and V2. When each black bar was presented separately (Fig. 1a, 'part 1' and 'part 2'), as expected, neurons did not respond because the parts were outside the receptive field of the neuron (red circle shows an example receptive field; responses are simulated). However, when the two parts were put together in the correct configuration (Fig. 1a, 'whole'), a percept of a white rectangle on top of the black ones emerged and a robust neural firing in V2 was observed, even though the parts by themselves or when presented in a wrong configuration caused little neural responses. In fact, the firing pattern to the illusory rectangle was comparable (albeit clearly weaker) to the neural response to the physically presented rectangle (Fig. 1a, 'actual rectangle'), even though no physical edges ever fell on the receptive fields of the recorded neurons.

Similar effects have been observed in many other paradigms. At the most basic level, Heeger (1992) and later Carandini and Heeger (2012) proposed divisive normalization as a canonical computation occurring at multiple levels throughout the brain where neural responses are normalized by their sum activity, leading to neural response saturation at high stimulus intensity levels (which is nonlinear). In the visual cortex, Brincat and Connor (2004) demonstrated that simple contour fragments (line segments and curves) are integrated into more complex shapes in inferotemporal (IT) cortex using both linear and nonlinear mechanisms. Similarly, in the domain of learning and expertise, Baker et al. (2002)

reported the emergence of a nonlinear part summation upon the repeated exposure of configurations of these parts, implicating mechanisms of holistic processing in IT.

To understand how such nonlinearities emerge in the brain, ideally we would like to be able to record in multiple brain areas. If the measured effects are coupled with behavioral measurements, we additionally would prefer a noninvasive technique that can be applied in human participants. Few tools available to neuroscientists have the capacity to cover large brain volumes and the sensitivity necessary to investigate the properties of neural representations which would indicate the presence of nonlinearities. In this paper, we discuss how multi-voxel pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data could be used for investigating nonlinear effects in the human visual system.

2. How MVPA works

Traditionally, fMRI has been used to compare the overall response strength to multiple stimulus conditions in one or several areas in the brain. For example, faces have been shown to activate the fusiform face area (FFA) more than any other stimulus category (Kanwisher et al., 1997). However, once a number of selective areas had been mapped, it became increasingly clear that a more powerful strategy would be necessary to understand processing within these areas at a finer scale. For instance, we might want to know whether identity is encoded in the FFA or we may ask whether there is a difference in encoding a particular stimulus feature in the early visual areas as compared to the higher visual areas. In the classical univariate analysis, the signal is averaged across all voxels within each region of interest (ROI), providing a measurement of the mean response activation per region of interest (ROI). Thus, comparing responses to different stimuli is limited to comparing single numbers. Such strategy is certainly not very robust – fine-grain differences might easily be washed out by noise, differences in activation level in each voxel, and variation across scans or participants.

Over the past decade, MVPA has become an established solution to this problem. Each ROI is composed of a number of voxels, each of them responding to a certain extent to a given stimulus. These voxel responses are thought to reflect the average neuronal and synaptic activity (Logothetis & Wandell, 2004). Each voxel responds in a particular way to one condition and in a different way to another one, giving rise to distinct patterns of response that remain roughly similar across multiple scans (because the brain responds similarly to the same stimulus). Armed with these patterns of response, we have not a single number but an *n*-dimensional vector of numbers per condition, where *n* is the number of voxels in that brain region. In order to reveal differences or similarities between the conditions, a simple correlation of these vectors might already suffice (Haxby et al., 2001; Misaki et al., 2010). If the ROI processes two conditions in a similar manner, a correlation of their responses is expected to be relatively high. Conversely, a lower correlation implies that the ROI is processing the two conditions differently.

A correlation is the simplest and most direct instance of MVPA. Another very popular and robust technique that has roots in machine learning literature is known as a linear Support Vector Machine (SVM; Kamitani & Tong, 2005; Op de Beeck et al., 2008; also see Misaki et al. (2010) for a thorough

comparison of these and other methods in the context of fMRI research). In a nutshell, using multiple samples of each condition (coming from the multiple scans of a participant), an SVM learns the optimal separating boundary in voxel space (called a hyperplane) between the conditions. Once trained, the SVM can be used on a fresh set of data (the testing set) as an independent means of quantifying how distinct the two conditions are. If the two conditions are very similar in that ROI, the learned boundary does not separate them well and, consequently, makes many mistakes. Thus, in the testing phase it produces a poor performance in classifying new data, with performance close to 50 % (i.e., chance level in a binary decision, meaning it can hardly distinguish between the two conditions). Highly dissimilar conditions, however, lead to a better than chance classification performance (the exact numbers can vary a lot, depending on the region of interest, stimulus conditions, task, size of dataset, and so on).

Taken together, MVPA provides a more sensitive analysis by utilizing more data points. In the next section, we show how it can be applied to investigate the linearity of processing in the brain, using several recently published datasets.

3. Detecting nonlinear transformations by MVPA with synthetic patterns

In this section, we discuss a paper by Baeck and colleagues (2013) to illustrate how MVPA can be used in the context of linear and nonlinear summation of elements. Baeck and colleagues (2013) investigated the representation of object pairs in the ventral visual pathway. Under normal circumstances, a pair of objects is expected to be represented as the linear combination of the representation of the single objects (Agam et al., 2010; Kaiser et al., 2014; MacEvoy & Epstein, 2009, 2011; Reddy et al., 2009; Zoccolan et al., 2005). However, if the two objects are placed in a familiar action configuration (e.g., a cork screw on top of a bottle of wine), their interaction becomes important to the visual system and is encoded above and beyond the two objects separately (just like the parts forming an illusory rectangle in the Kanizsa display in Fig. 1a), as manifested in behavioral, fMRI, and TMS measures (Green & Hummel, 2006; Kim & Biederman, 2011; Kim et al., 2011; Riddoch et al., 2003, 2006). In other words, these action pairs are represented as a “whole” that is not equivalent to a mere sum (or mean) of the two objects in the pair.

To shed light on the neural mechanisms governing these effects, Baeck et al. (2013) compared representations of familiar action pairs to several control conditions: familiar non-action pairs where the two objects were in the wrong configuration (e.g., a cork screw below a bottle of wine), non-familiar action pairs (e.g., a cork screw on top of a piece of paper), and non-familiar non-action pairs (e.g., a cork screw below a piece of paper). Baeck et al. (2013) presented participants with these different configurations of two objects as well as with each object separately (either above or below the fixation spot) and recorded fMRI responses in the shape-selective posterior and anterior parts of the lateral occipital cortex (pLOC and aLOC; however, in this article, we refer to them LO and pFs for consistency across different studies, although the precise localization details may differ). As a result, they obtained brain responses to each stimulus separately and to their combinations. Then, employing and extending an analysis developed by MacEvoy and Epstein (2009), they investigated what combination of the

representations of the two objects presented in isolation would best explain the observed response to the object pair.

Several options are popular in the literature. First, the neural response to the pair of objects could be simply a sum of the responses to each object separately, taking the famous Gestalt motto literally. This idea, however, is somewhat simplistic when neural mechanisms are taken into consideration. If responses were adding up linearly, with several more objects a maximal neural firing rate would be reached and such summation would quickly break down. Consistent with this explanation, Zoccolan et al. (2005) found no evidence for summation in monkey IT. Rather, they found that most neurons perform a mean operation, where the neural response to the pair would be best described be a simple mean of the responses to each object separately (see their Fig. 3). This simple combination of responses has also been observed at larger scales in several fMRI studies (Kaiser et al., 2014; MacEvoy & Epstein, 2009, 2011) where the mean is computed between patterns of responses, i.e., between voxels.

In a later study, however, Zoccolan and colleagues (2007) extended their previous findings. They reported that while neurons that are very selective to one object tend to compute a simple mean when a second object is also presented, a majority of neurons in IT perform a weighted mean where the object that by itself elicits a stronger response receives a higher weight (see their Fig. 8). In the most extreme cases, where neurons do not show a clear selectivity pattern, Zoccolan et al. (2007) found that a max operation is performed, such that the response of a pair of objects is equal to the response of a single object that activates the neuron more.

Note that all three cases (mean, weighted mean, and maximum) are in fact nonlinear operations that can be expressed mathematically as $c_{max} \max(x_1, x_2) + c_{min} \min(x_1, x_2) + c_{offset}$, with $c_{max} = c_{min} = 0.5$ in the case of the mean, $c_{max} = 1$ and $c_{min} = 0$ for the maximum operation, and all other cases where $c_{max} \neq c_{min}$ constituting a weighted mean. In other words, the summed response to objects is always some sort of a weighted mean of their original responses, also influenced by the number of objects, leading even the simple mean to be nonlinear. The definition of weighted mean is not consistent throughout the literature though, with some authors considering a linear version of a weighted mean where an object at a particular location, for example, always receives a higher weight than the other location (Gawne & Martin, 2002). However, most studies use the definition of a weighted mean as described above.

Although all three cases are instances of the weighted mean, we want to emphasize that they lead to different interpretations regarding the underlying computations. A simple mean is the easiest mechanism of combining neural responses, because no preference is given to any one object. In the case of a weighted mean, a more salient object is always weighted more in the combination. In contrast, with the max operation, no information about the less salient object is retained in the combination. Our interest is in understanding where in the continuum of the parameter space of a single model (the weighted mean) the optimal parameters lie: close to the mean, to the max, or somewhere in between. Consequently, a max operator is often regarded as a more “nonlinear” function than a weighted mean, which, in turn, is more “nonlinear” than a simple mean where the only nonlinearity is normalization by the number of objects. We adopt such ordering of response “nonlinearity” in this paper as well, and because of this qualitative difference many studies focus on distinguishing between these three

alternatives. Importantly, the weighted mean model is still rather linear as compared to other possibilities, such as the illusory rectangle example (Fig. 1a) by von der Heydt et al. (1984), where the pair could be represented in a way that is not related at all to the representations of the single objects. However, such nonlinear functions are often more idiosyncratic and not suited for studies of the generic nonlinear mechanisms in the brain. We consider how to deal with such functions in Section "Using MVPA when the whole cannot be easily predicted from parts".

In the case of Baeck et al. (2013), for unfamiliar action pairs or pairs in the wrong configuration (non-action pairs), the observed response should be the mean of the responses to each basic stimulus independently, as found in similar study by MacEvoy and Epstein (2009), or at least closer to the simple mean than for familiar action pairs where one would expect a more pronounced weighted mean, reflecting the interaction between the objects, or perhaps even a max operation, such that only the maximally preferred object dominates the response.

This description of expected outcomes easily translates into an MVPA analysis. In particular, we can produce synthetic patterns of response to the two objects presented as a pair by computing the mean or applying a max operator (or any other function of our choice) of the two patterns of responses as recorded for each object separately (Fig. 2a). To understand how synthetic patterns can be used, consider first what this procedure would yield in a single voxel (Fig. 2b). We can plot the synthetic responses against the actual responses. The closer a particular synthetic pattern is to the actual responses (a perfect correlation is indicated with a dashed line), the better this model of combining responses is. In Fig. 2b, for example, we see that the weighted mean provides the best fit while the other two operators (simple mean and max) appear to deviate more from the actual data.

To evaluate how well these synthetic patterns capture the full patterns of response, we can train an SVM classifier to distinguish between various synthetic pairs and test it on the actual measured responses to the presented pairs. This performance is further compared to the decoding performance when the classifier is both trained and tested on the actual responses to the presented pairs of objects. The closer the decoding performance using synthetic training pairs matches the decoding performance of the actual data, the better the function used to create the synthetic pair describes the relationship between the responses to the single objects and the responses to the actual object pairs (Fig. 2c).

Baeck et al. (2013) evaluated four alternatives: mean, weighted mean, max, and min. The mean, max, and min can all be computed immediately, while for the weighted mean, the weighting parameters need to be computed first by fitting the general mean model to the data (for details, see Section "Quantifying nonlinearities"). They found that responses synthesized using a weighted mean (with $c_{max} \neq c_{min}$) accounted best for their results (Fig. 3, 'other configurations' bars). Notice that this is not a trivial outcome of the weighted mean model being more general than the other three alternatives. They could have found that the optimal parameters of the general weighted mean model were close to the mean ($c_{max} = c_{min} = 0.5$), for example. Instead, the authors showed that the found parameters c_{max} and c_{min} differed reliably from each other and from zero, meaning that the optimal weights genuinely reflected the weighted mean and not any other option. The difference between the weighted mean and a simple mean was not large, but consistent across subjects and as such significant. Moreover, the strongest

single-object response was weighted more than the weakest single-object response in each voxel. Interestingly, this observation held true for all conditions irrespective of whether they were familiar action pairs or not (Fig. 3, 'familiar action pair' bars), meaning that familiar actions did not elicit nonlinear addition and are not represented in a more special manner than any random pair of objects.

The authors further tested whether this finding would hold when the configuration is relevant to the task the participants are performing. To test this, participants were asked to judge how well the object pairs were positioned to perform an action together. Under these circumstances, the difference in weights between the max and min response to the single objects increased for the familiar action pairs, but not for all other types of configurations. The effect of task context suggests the contribution of higher areas (beyond LO and pFs) in producing the behavioral familiar action pair effect. However, even though a difference in the weights between the different types of object pairs was found, in both cases the relationship between the single objects and the object pairs was best described by a weighted average.

This finding complements similar results using other techniques. For example, using electroencephalography (EEG), Agam et al. (2010) reported a weighted mean encoding of random object pairs. In a neurophysiological study, as noted above, Zoccolan et al. (2007) reported a whole spectrum of combination functions in the visual cortex. The most selective neurons tended to perform a simple averaging, while the noisiest units performed a max operation. Across multiple neurons (i.e., what is measured in fMRI), this relationship translates into a weighted average, consistent with findings by Baeck et al. (2013).

4. Quantifying nonlinearities

While in the previous section we described the idea of using MVPA as a means to figure out which operation best describes the actual combination of voxel responses in the brain, we still have little understanding how these effects could be achieved, or, in other words, we have not yet quantified the amount of nonlinearity in this data. How could we do that? Below we present three possible approaches.

The simplest approach that works for a weighted mean is to try several combinations of parameters (e.g., .5 and .5, .6 and .4, .7 and .3, etc.) and choose the one that leads to the best decoding performance (or interpolate to find the best estimate).

Alternatively, we can estimate these parameters directly by conducting a more elaborate analysis, as suggested by MacEvoy and Epstein (2009) and also employed by Baeck et al. (2013). This analysis consists of three steps: searchlight, regression, and combining the two. In the first step, a searchlight is conducted where for each voxel within an ROI, a cluster of all other voxels within a small radius (e.g., 5 mm) is defined and an SVM classification for the 'whole' stimuli is performed on these voxels, as usual. This procedure amounts to finding clusters that contain the most informative voxels for our classification within an ROI (we call the resulting ranking the "cluster classification rank"). Next, we perform a linear regression between 'parts' and 'whole' stimuli for each voxel using the generic

weighted mean model as described above. Finally, parameters from this regression (R^2, slope or coefficient weights, and intercept) are related to the cluster classification rank (Fig. 4). The idea here is that the most informative clusters (data points on the right side of the panels in Fig. 4) should be the closest to the "true" relationship between responses to pairs and their components. As can be seen in the data from Baeck et al. (2013) in the top left panel of Fig. 4, this assumption holds since R^2 of the weighted mean model increases with better classification rank.

By averaging regression parameters from several top ranked clusters (e.g., 20 clusters in Baeck et al., 2013), we finally can quantify the weighted mean parameters and see if they are close to the simple mean, maximum, or lie somewhere in between. For example, the intercept (Fig. 4, top right) converged to a value statistically not different from zero, meaning that there was no additional constant neural activity for pairs of objects. Moreover, the weighted mean regression employed two regressors, the max and the min. As seen in Fig. 4, the value of max coefficient increased with an increasing cluster classification rank (Fig. 4, bottom left), while there was no such relationship with the min coefficient (Fig. 4, bottom right), and the mean values of the max and min coefficients differed reliably (0.66 versus 0.28, $p < 0.001$), thus confirming that this was a weighted mean and not a simple mean operation. Also, since the min value was not zero, it also differed from a pure max operation. Note, however, that this powerful analysis is only viable if the dataset is good enough (sufficient number of runs, good scanner signal-to-noise parameters, little noise and so on). This is the case in the data of Baeck et al (2013), in which SVM discrimination between object pairs was around 90% in LOC.

5. Using MVPA when the whole cannot be easily predicted from the parts

In the previous sections, we demonstrated how synthetic patterns of response could be used to assess the type of relationship between responses to stimuli presented separately and the responses to a combination of the single objects ('whole' stimulus). This method has several disadvantages, however. A researcher needs to present each part separately and all their combinations in order to be able to produce synthetic patterns of response. Consequently, the number of conditions in the experiment quickly increases and it may not be feasible to obtain sufficient data in a single or even two fMRI sessions. Moreover, these conditions are not equivalent (the number of objects on the screen differs), which may cause certain attentional or saccadic confounds, or make it difficult for the researcher to come up with a single engaging task for all conditions. Additionally, the decoding performance for the 'whole' stimulus needs to be sufficiently high. Decoding performances using synthetic patterns are more noisy compared to the decoding of actual presented stimuli. In case the signal is not clear enough, it is very likely that none of the decoding performances using synthetic patterns are above chance level, making it impossible to choose the 'best approximation'.

Finally, by design, the analyses proposed in the previous section can only capture to what extent parts contribute to the final percept. However, the whole that emerges (e.g., the illusory white rectangle in Fig. 1a) cannot be accounted for by any combination of the two parts in this framework, leading to poor fits of the weighted mean model to the data. In those instances, one possibility would be to use

models that are explicitly aimed at producing these wholes. For example, in the case of Fig. 1a, several computational models have been developed to produce these illusory shapes(e.g., Grossberg & Mingolla, 1985; Kogo et al., 2010). There are also other general models of the ventral visual pathway aimed at explaining object recognition (e.g., Serre et al., 2007; Yamins et al., 2014). So now, instead of asking which function describes our data better, we can use MVPA to compare competing models of vision. First, we provide each model with parts and wholes displays and read out their responses. Next, just like in Baeck et al. (2013), synthetic patterns of response are generated from responses to parts, and an MVPA is applied both to the simulated and the actual neural data. The best model is the one whose MVPA outcomes match the actual neural data best.

This approach was taken by Zoccolan and colleagues (2007), where the observed trade-off between neuronal selectivity and amount of nonlinearity was compared to the outputs of two models of vision, a simple toy model by Riesenhuber & Poggio (1999) and a more evolved version of it that has been reported to match human performance in a certain object recognition task (Serre et al., 2007). The authors found that both models could account for their data comparably well, which is perhaps not so surprising given that both models are based on the same HMAX architecture. Employing other, more distinct models might in fact show larger discrepancies.

However, in many cases using these models is not possible. Moreover, one has to be very careful about comparing different models because they may have a different number of free parameters or one might generalize the other, and so forth. It would therefore be desirable to have a method to investigate nonlinearities without knowing or restricting oneself to a particular combination function. When the research question revolves around getting insight whether a combination of parts leads to a qualitatively different whole, e.g. one where parts representations are lost or lead to a qualitatively different percept (as in the case of Fig. 1), one can use a simpler technique (Fig 2d, “wholes dominate” model). We illustrate it in this section with a paper by Kubilius and colleagues (2011). Similarly to Baeck et al. (2013), Kubilius and colleagues (2011) investigated what happens to parts when they are incorporated in wholes. However, to maximize chances of finding a highly nonlinear operation, , they employed a robust Gestalt phenomenon, known as the configural superiority effect described by Pomerantz and colleagues (1977; see also Pomerantz & Portillo, 2011). Consider the display shown in Fig. 5a, orange. It consists of four line segments in four quadrants, three of them in one orientation and the remaining one differing in its orientation by 90°. The task for an observer is to find the quadrant where the oddly oriented line segment is. It is not a difficult task but it may still take you a moment to correctly respond that it is located at the top left. Now consider adding a corner, as shown in Fig. 5a, middle, to all four of these line segments. Notice that in fact we are not changing much because the exact same information is added to all four elements. So, it is completely redundant and it could actually be thought of as adding noise or clutter to the image, which, if anything, should make the observer’s task only more difficult. However, the result is in fact the opposite – it is now trivial to detect the odd element because these parts are now combined into wholes with emergent, configural properties: a triangle that is popping out from the three arrows (Fig. 5a, dark red displays).

This is a powerful effect with participants showing a couple of hundred milliseconds difference between the two conditions. We can therefore take this effect into the MRI scanner and investigate

whether such nonlinear combination of parts can be observed neurally as well. However, unlike Baeck et al. (2013), this study used only the four "parts" displays (only line segments present with the odd one in the different quadrants) and the four "wholes" displays (triangles and arrows present), omitting the corners condition. In this instance, this design choice was motivated by the desire to have participants perform the task while being scanned. With the corners condition included, the study would have had to resort to passive viewing or an orthogonal task, thus risking to diminish the strength of the configural superiority effect.

In order to investigate where in the visual system the combination of a line segment and a corner into a triangle or an arrow preserves part representations and where the whole dominates instead, the following MVPA strategy was employed. First, the classification performance was estimated of discriminating only between parts displays and only between wholes displays (six possible pairwise discriminations for each). If the added corner in the stimulus is combined in a general weighted mean manner, decoding of the parts displays and decoding of the wholes displays should be similar because the same non-informative corner is added to all four stimuli (Fig. 2d, "parts dominate") or even slightly worse because adding a corner amounts to adding more noise. However, if adding a corner drastically changes the neural representations of the resulting wholes such that they no longer resemble the neural representations of the parts, as expected from the behaviorally observed effect, decoding of wholes would be better than decoding of parts (Fig. 2d, "wholes dominate"). In other words, based on behavioral responses, according to which triangles appear more dissimilar to arrows than line segments of different orientation, we expect that the neural representations of triangles versus arrows might also be more dissimilar. Moreover, we can train a classifier to distinguish between wholes and then test it on the parts (or vice-versa). For part-dominated combinations, decoding of parts from wholes should remain robust, but for heavily whole-dominated combinations, decoding of parts should be worse if not at chance level.

Importantly, both hypotheses can co-exist in the network of regions in the visual cortex, with different regions performing different computations. Therefore, a number of visual areas were identified in the ventral visual cortex, namely, early visual areas V1, V2, and V3, and higher visual areas LO and pFs, using an independent localizer. The resulting representations of triangles and arrows in the early visual cortex were closer to the simple combinations of the line segment and a corner (Fig. 5b). In particular, decoding of parts in the early visual cortex was better than decoding wholes. Conversely, decoding in higher visual areas was better for wholes than for parts, indicating that towards the end of the visual shape processing stream, increasingly nonlinear transformation of parts occurs. In fact, in the most anterior region investigated, namely, pFs, parts could no longer be decoded from whole shapes. So unlike Baeck et al. (2013), in this case vastly nonlinear summation effects where part representations are lost were found in the lateral occipital cortex, in line with a number of studies using other techniques (e.g., Baker et al., 2002; Brincat & Connor, 2004).

To more directly compare their results with these results from Kubilius et al. (2011), Baeck et al. (2013) analyzed a subset of the data from the latter study in the same way as the former study. In the study of Baeck et al. (2013), we can also compare the performance on decoding single objects with the performance on decoding objects in pairs in which only one object is different (Fig. 5b). Here we have

the same situation central to the design of Kubilius et al. (2011), namely that we compare the same stimulus display without (single object) or with (pair) the addition of a context which is not informative at all (as it is the same in all conditions which have to be discriminated by the decoder). In LOC, decoding performance was higher in the single-object condition compared to the performance for pairs with one object in common in the same position. This is the same result as found in retinotopic visual areas but not in LOC by Kubilius et al. (2011): better performance for the parts than the wholes condition. In other words, while the combination of parts into wholes was captured fairly successfully with the weighted mean model by Baeck et al. (2013), in other instances we found evidence for an even more complex part transformation where the whole rather than the parts is enhanced.

Furthermore, note that this result alone and MVPA in general cannot provide insights into causal relations of the observed effect. In this study, for example, a behavioral pop-out effect is supported by the response in the LOC, yet we do not know whether LOC is causally involved in producing it or whether it is an epiphenomenal activation that we can decode but plays no role in the behavioral phenomenon. De-Wit, Kubilius and colleagues (2013) resolved this question by testing neuropsychological patient DF who lacks LOC and, as a result, is impaired in many (but not all) visual recognition tasks (Bridge et al., 2013; Goodale & Milner, 1992). The reasoning was that if LOC is truly responsible for the observed nonlinear behavioral effect, patient DF should find the task of detecting triangles among arrows just as challenging as the one with line segments, even if it is hard for us to imagine such an effect. Remarkably, when tested behaviorally, not only did DF show difficulties in performing the task with triangles and arrows but also her performance was worse than with line segments (Fig. 5b), surprisingly in line with decoding performance in early visual areas reported in Kubilius et al. (2011). Since DF is mainly relying on early visual cortex in performing this task, we asked if her performance could be accounted for by a simple computational model that reflects some of the key features of V1, namely, units preferentially responding to oriented bars of several spatial frequencies and multiple locations. This simulation supported the pattern of results observed in DF and in the early visual cortex in the study by Kubilius et al. (2011) with the model differentiating better between parts than between wholes displays. Taking these studies together, we see that by combining multiple techniques it is sometimes possible to distinguish the roles of several regions in the ventral visual stream in producing nonlinear effects.

Moreover, these studies imply that at least in the case of the configural superiority effect, a feedforward emergence of this phenomenon is possible, with early visual areas involved in a simple part-based processing and later areas producing heavily nonlinear representations dominated by whole shapes rather than their constituent elements. But not all Gestalts are equal. Consider, for example, a moving diamond display (Fang et al., 2008) where the diamond shape is partially occluded by three vertical rectangles of the same color as the background, effectively creating the display of four independently moving line segments. Under certain luminance and speed conditions, however, it is possible to perceive either the four line segments moving independently (parts condition) or the diamond as a whole (whole condition). Using this setup, both Fang et al. (2008) and de-Wit and colleagues (2012) reported that when a diamond is perceived, responses in V1 are decreased as compared to when the four separate lines are reported. Based on the known properties of V1 (e.g.,

small receptive field sizes, tuning to oriented bars but not their combinations), it would be difficult to come up with an explanation how V1 by itself could produce such effects. Instead, the authors of these studies argued that the presence or absence of the whole diamond shape is communicated to V1 in a top-down fashion. Indeed, this idea was supported by their findings in LO that exhibited an increase in activation when the whole diamond was perceived. Thus, comparing the Kubilius et al. (2011) study to Fang et al. (2008) and de-Wit et al. (2012) provides hints about potential differences between processes when parts are combined into wholes. In particular, just like in the case of Baeck et al. (2013), in some cases later processing stages and feedback from them might be crucial in forming the holistic percept.

6. Conclusion

In this paper, we discussed how fMRI MVPA analysis could be employed in understanding nonlinear effects in the visual system. We demonstrated that MVPA could be used to compare real and synthetic response patterns of configural displays based on chosen combinations of voxel responses to their constituent parts, such as simple or weighted mean, maximum, minimum, or any other nonlinear function. In order to not only describe but also quantify these effects, we suggested several analyses, one involving a combination of searchlight, MVPA, and regression analyses, and another using computational models of visual processing. We also showed a simpler method of unveiling nonlinear effects based on comparing voxel responses to parts and to wholes. Taken together, MVPA stands as a powerful technique to investigate nonlinear both locally (within an area) and across a network of regions.

Acknowledgements

This work was supported in part by a Methusalem Grant (METH/08/02) awarded to Johan Wagemans from the Flemish Government, the IUAP P7/11 grant from the Federal research action, Belgium and ERC-2011-Stg-284101 grant from the European Research Council awarded to Hans P. Op de Beeck. Jonas Kubilius and Annelies Baeck are research assistants of the Research Foundation—Flanders (FWO).

Statement of Open Science. Figures in this paper have been produced using free and open source tools, including Python, matplotlib, seaborn, and Inkscape.

References

Agam, Y., Liu, H., Papanastassiou, A., Buia, C., Golby, A. J., Madsen, J. R., & Kreiman, G. (2010). Robust selectivity to two-object images in human visual cortex. *Current Biology*, *20*(9), 872–879. doi:10.1016/j.cub.2010.03.050

Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2013). The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule. *NeuroImage*, *70*, 37–47. doi:10.1016/j.neuroimage.2012.12.023

Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, *5*(11), 1210–1216. doi:10.1038/nn960

Bridge, H., Thomas, O. M., Minini, L., Cavina-Pratesi, C., Milner, A. D., & Parker, A. J. (2013). Structural and functional changes across the visual cortex of a patient with visual form agnosia. *The Journal of Neuroscience*, *33*(31), 12779–12791. doi:10.1523/JNEUROSCI.4853-12.2013

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, *7*(8), 880–886. doi:10.1038/nn1278

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. doi:10.1038/nrn3136

de-Wit, L. H., Kubilius, J., Op de Beeck, H. P., & Wagemans, J. (2013). Configural Gestalts remain nothing more than the sum of their parts in visual agnosia. *I-Perception*, *4*(8), 493–497. doi:10.1068/i0613rep

de-Wit, L. H., Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2012). Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. *Journal of Vision*, *12*(11). doi:10.1167/12.11.12

Fang, F., Kersten, D., & Murray, S. O. (2008). Perceptual grouping and inverse fMRI activity patterns in human visual cortex. *Journal of Vision*, *8*(7), 2. doi:10.1167/8.7.2

Gawne, T. J., & Martin, J. M. (2002). Responses of primate visual cortical neurons to stimuli presented by flash, saccade, blink, and external darkening. *Journal of Neurophysiology*, *88*(5), 2178–2186. doi:10.1152/jn.00151.200

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. doi:10.1016/0166-2236(92)90344-8

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1107–1119. doi:10.1037/0096-1523.32.5.1107

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, *92*(2), 173–211. doi:10.1037/0033-295X.92.2.173

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. doi:10.1126/science.1063736

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(02), 181–197. doi:10.1017/S0952523800009640

Kaiser, D., Strnad, L., Seidl, K. N., Kastner, S., & Peelen, M. V. (2014). Whole person-evoked fMRI activity patterns in human fusiform gyrus are accurately modeled by a linear combination of face- and body-evoked activity patterns. *Journal of Neurophysiology*, *111*(1), 82–90. doi:10.1152/jn.00371.2013

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. doi:10.1038/nn1444

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex*, *21*(8), 1738–1746. doi:10.1093/cercor/bhq240

Kim, J. G., Biederman, I., & Juan, C.-H. (2011). The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: A transcranial magnetic stimulation study. *The Journal of Neuroscience*, *31*(22), 8320–8324. doi:10.1523/JNEUROSCI.6450-10.2011

Kogo, N., Strecha, C., Van Gool, L., & Wagemans, J. (2010). Surface construction by a 2-D differentiation–integration process: A neurocomputational model for perceived border ownership, depth, and lightness in Kanizsa figures. *Psychological Review*, *117*(2), 406–439. doi:10.1037/a0019076

Kubilius, J. (2014, July 24). Le FishBird. Retrieved from http://figshare.com/articles/Le_FishBird/1116281

Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2011). Emergence of perceptual Gestalts in the human visual cortex: The case of the configural-superiority effect. *Psychological Science*, *22*(10), 1296–1303. doi:10.1177/0956797611417000

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD Signal. *Annual Review of Physiology*, *66*(1), 735–769. doi:10.1146/annurev.physiol.66.082602.092845

MacEvoy, S. P., & Epstein, R. A. (2009). Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Current Biology*, *19*(11), 943–947. doi:10.1016/j.cub.2009.04.020

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. doi:10.1038/nn.2903

McKone, E., Aimola Davies, A., Darke, H., Crookes, K., Wickramariyaratne, T., Zappia, S., … Fernando, D. (2013). Importance of the inverted control in measuring holistic face processing with the composite effect and part-whole effect. *Frontiers in Psychology*, *4*, 33. doi:10.3389/fpsyg.2013.00033

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. doi:10.1016/j.neuroimage.2010.05.051

Op de Beeck, H. P., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *The Journal of Neuroscience*, *28*(40), 10111–23. doi:10.1523/JNEUROSCI.2511-08.2008

Pomerantz, J. R., & Portillo, M. C. (2011). Grouping and emergent features in vision: Toward a theory of basic Gestalts. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1331–1349. doi:10.1037/a0024330

Pomerantz, J. R., Sager, L. C., & Stoever, R. J. (1977). Perception of wholes and of their component parts: some configural superiority effects. *Journal of Experimental Psychology. Human Perception and Performance*, *3*(3), 422–35.

Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences of the USA*, *106*(50), 21447–21452. doi:10.1073/pnas.0907330106

Riddoch, M. J., Humphreys, G. W., Edwards, S., Baker, T., & Willson, K. (2003). Seeing the action: neuropsychological evidence for action-based effects on object selection. *Nature Neuroscience*, *6*(1), 82–89. doi:10.1038/nn984

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. doi:10.1038/14819

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the USA*, *104*(15), 6424–6429. doi:10.1073/pnas.0700622104

Transfer Printables - Bird Silhouettes - Swallows. (n.d.). Retrieved from http://thegraphicsfairy.com/transfer-printables-bird-silhouettes-swallows/

Von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, *224*(4654), 1260–1262. doi:10.1126/science.6539501

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172–1217. doi:10.1037/a0029333

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the USA*, *111*(23), 8619–8624. doi:10.1073/pnas.1403112111

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747 – 759. doi:10.1068/p160747

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *The Journal of Neuroscience*, *25*(36), 8150–8164. doi:10.1523/JNEUROSCI.2058-05.2005

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *The Journal of Neuroscience*, *27*(45), 12292–12307. doi:10.1523/JNEUROSCI.1897-07.2007

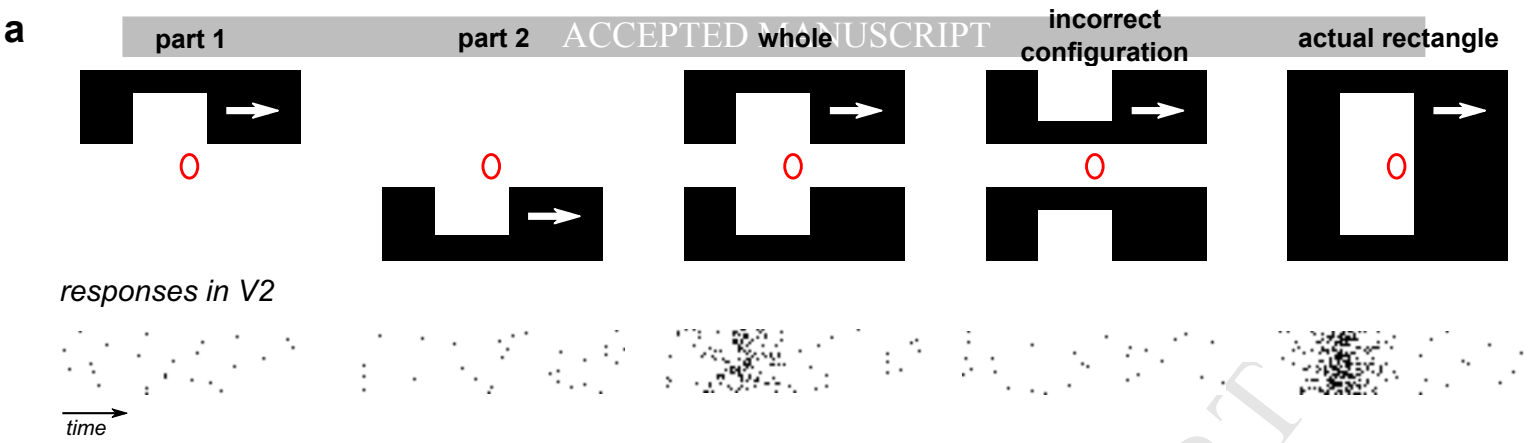
Figure 1. Examples of nonlinear part combination. (a) A modified Kanizsa display, as used by von der Heydt et al. (1984), and simulated neural responses of a neuron with its receptive field marked in red. Direction of the arrow indicates motion of the white notches. (b) An ambiguous figure-ground display, similar to the famous drawings by M.C. Escher. Although many people would see fish in the water and birds in the sky, in fact both animals are present in the water and in the sky but the color of background defines what is seen as a figure. (Figure taken from Kubilius (2014) as permitted by the Creative Commons Attribution License.) (c) Composite face effect. The top half of the face is Barack Obama and the bottom half is Will Smith. When the two halves are properly aligned, it is hard to tell the identity of each half. (Figure taken from McKone et al. (2013) as permitted by the Creative Commons Attribution License.)

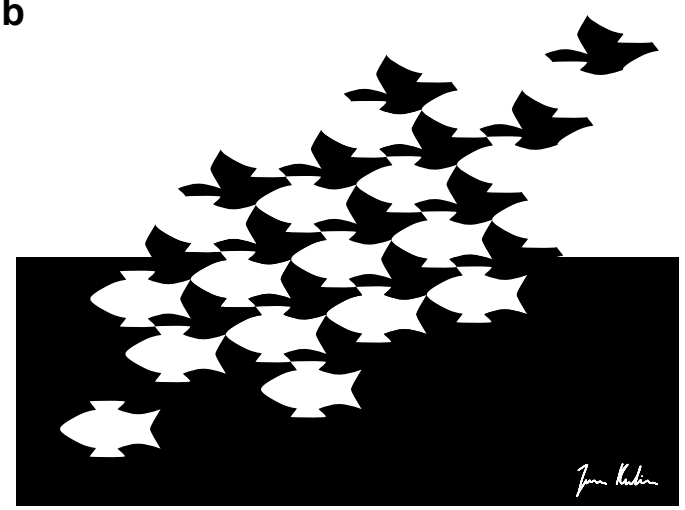
Figure 2. Using MVPA for investigating nonlinear processes in the brain. (a) Example fMRI responses to two parts and their (nonlinear) combination ("whole"; only nine voxels shown). Responses to parts are then synthetically combined using a mean, weighted mean, or max operation. (b) A plot of synthetic responses to pairs (two parts) against the actual brain response to the whole (each dot indicates a voxel response to a particular exemplar or trial). In this example, the weighted mean appears to fit the actual response best. (c) MVPA decoding performance using these synthetic patterns in comparison to decoding of the actually recorded patterns ("whole" condition; see Section 3 "Detecting nonlinear transformation by MVPA with synthetic patterns"). (d) MVPA decoding without synthetic patterns (see Section 5 "Using MVPA when the whole cannot be easily predicted from the parts"). When decoding of parts is similar to or better than decoding wholes, neural processing is more likely preserve parts (closer to the weighted mean model). When this decoding is worse, it indicates that part representations are no longer prominent and that the whole dominates.

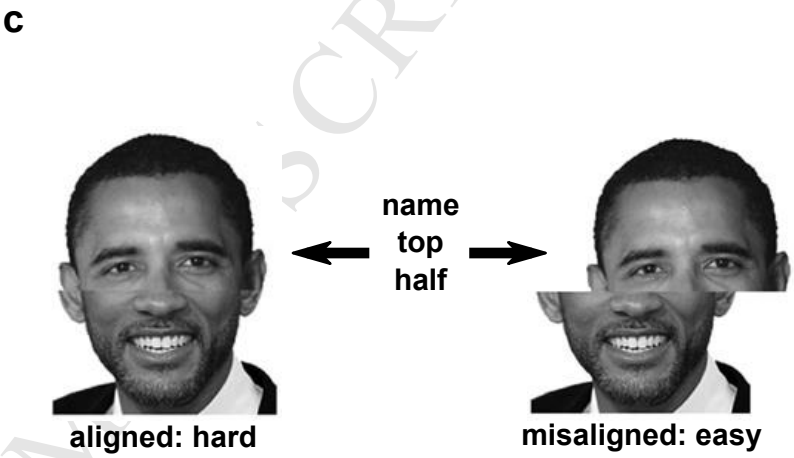
Figure 3. Decoding of familiar action pairs and other types of object configurations (Baeck et al., 2013). The weighted mean fits these data best as it is the closest to the decoding of the actual pair. Error bars indicate the standard error of the mean (s.e.m.) across participants.

Figure 4. Cluster classification ranking outcome for estimating weighted mean parameters with more informative clusters having a higher rank (data points on the right side of each figure) for data in Baeck et al. (2013).

Figure 5. Differences in decoding performance of parts and wholes in the visual cortex. (a, b) Stimuli and data from Kubilius et al. (2011) and for patient DF (de-Wit, Kubilius et al., 2013). (c, d) Data from Baeck et al. (2013) for comparable conditions, i.e., when the same object is added to the initial display. Notice that a combination of parts yields vastly different computations in LOC depending on stimuli. Whereas in Baeck et al. (2013) adding a noninformative object only spoiled the neural representations in LOC, in Kubilius et al. (2011) addition of a noninformative corner helped significantly in LOC but not in early visual areas.

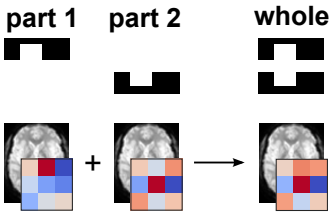




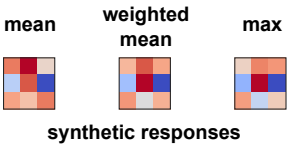


a

record fMRI patterns of response

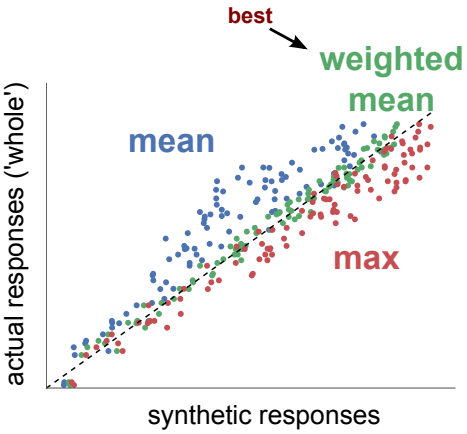


combine parts using:



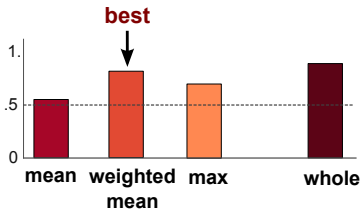
b

responses in a single voxel



c

multi-voxel: decoding synthetic vs. actual



d

multi-voxel: decoding parts vs. wholes